Title
Auto Insurance Tenure Prediction and Analysis

Author
Chen, Yi

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Auto Insurance Tenure Prediction and Analysis

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics

by

Yi Chen

2020

ABSTRACT OF THE THESIS

Auto Insurance Tenure Prediction and Analysis

by

Yi Chen
Master of Applied Statistics
University of California, Los Angeles, 2020
Professor Yingnian Wu, Chair

The purpose of this project is to understand the main factors that drive customer tenure within auto insurance industry for six or more years. The analysis is based on three years of the J.D. Power Auto Insurance survey data. For the analysis, multiple binary machine learning algorithms were implemented and measured to classify whether customers would stay with the same insurer for more than six years. Random forest was found to be the most robust model as compared to logistic regression, decision trees, and xgboost.

The thesis of Yi Chen is approved.

Frederic R Paik Schoenberg

Nicolas Christou

Yingnian Wu, Committee Chair

University of California, Los Angeles

2020

Table of Contents

1 Introduction 1

2 Data Preparation 2

2.1 Data Description 2

2.2 Handling Missing Data 2

2.3 Response Variable – Tenure 3

2.4 Variables Reduction – Correlation and Multicollinearity 4

2.5 Feature Engineering 6

2.5.1 Binarization 6

2.5.2 Seasonality 7

2.5.3 Interaction Effect 8

2.5.3 Data Descriptive 9

3 Binary Classifications Modeling 12

3.1 Data Standardization 12

3.2 Logistic Regression 12

3.2.1 Backward Selection 13

3.2.2 Compare Residual Deviance 14

3.2.3 Check Multi-collinearity 16

3.3 Logistic model performance 18

3.3.1 Assess Model Performance 18

3.3.2 Methods of Improving Model Accuracy 20

3.5 Decision Tree 24

3.6 Random forest 25

3.7 xgboost 27

3.8 Conclusion 28

4 Model Interpretation and Data Analysis 29

4.1 Models Interpretation - Logistic Regression 29

4.1.1 Binary Logistic Regression 29

4.1.2 Multi-class Logistic Regression 31

4.2 Models Interpretation - Decision trees 33

4.3 Data Analysis - Generational Difference 34

5 Further Research 37

6.1 Getting More Data 37

6.2 Further Data Analysis 37

6. Reference 38

Table of Figures

Figure 1: Histogram of Auto Tenure 3

Figure 2: Bart Chart for Response Variable – Tenuremorethan6years..... 4

Figure 3: Variables Correlations Graph..... 5

Figure 4: Boxplots of Credit Level and Generation..... 7

Figure 5: Stepwise Logistic Regression..... 14

Figure 6: Residual Deviance..... 15

Figure 7: VIF for Logistic Regression with All Variables 17

Figure 8: ROC Graph..... 20

Figure 9: K-fold Cross Validation Accuracy Scores 22

Figure 10: Eigen Values for PCA 23

Figure 11: PCA R Outputs..... 24

Figure 12: Complexity Parameter for Decision Tree..... 25

Figure 13: Out of Bag Error for Number of Trees..... 26

Figure 14: Out of Bag Error for Number of Variables 27

Figure 15: Feature Importance for Xgboost..... 28

Figure 16: Odds Ratio for Binary Logistic Regression 30

Figure 17: Odds Ratios for Multinominal Logistic Regression..... 33

Figure 18: Decision Trees..... 33

Figure 19: Variables Importance for Decision Tree 34

Figure 20: Decision Trees for Older Generation 35

Table of Tables

Table 1: Year Seasonality 8

Table 2: Confusion Metrics 19

Table 3: Summary of AUC and Accuracy at Different Level of Threshold 20

Table 4: Summary of All Methods Training and Testing Accuracy 28

CHAPTER 1

1 Introduction

What drives customer longer lifetime expectancy with the same insurer is a hot topic in the auto insurance industry. It is well known within the insurance industry that loyalty, lower premium, and better claims services are key factors drive customer retention. What is less understood are the hidden factors that contribute to customer longer tenure. How generation effects tenure is also not well understood. To better understand these factors, J.D. Power Auto Insurance survey data from 2017 through 2019 was investigated.

In chapter 2, details on data cleaning, data transforming, and selection of key independent variables for the modeling are discussed. The analysis is a combination of statistical methods and business acumen. In chapter 3, customer retention for six or more years is predict based on a few binary machine learning methods such as logistic regression, decision tree, random forest and Xgboost. Model accuracy of each algorithm was compared and recommendations are given on which is the most robust model for classifying customer retention. In chapter 4, modeling results are explored to better elucidate customer tenure for more than 6 years. In chapter 5, suggestions on how to improve on and advance the research are given.

CHAPTER 2

2 Data Preparation

In this chapter, I will explore the variables relationships with the response variable tenure and conduct the feature engineering exercise to select the key variables that will be used for modeling later.

2.1 Data Description

The data is based on the J.D. Power Auto Insurance survey data from 2017 through 2019. The initial dataset consists of 100,000 + rows with 1000+ variables.

While removing un-related variables with the response variable tenure, we still have 100+ potential exploratory variables. Running models with large set of variables is not wise as we are likely to see the multicollinearity issue due to correlations within variables. Therefore, the exploratory analyses are performed to select the important variables to be used for modeling.

2.2 Handling Missing Data

Since some of the questions were not applied to certain groups in a survey, we will be required to dealing with some missing data. When the data is randomly missing, we looked at the missing rate%. We can either remove the missing values or impute the missing values with mean. Unfortunately, removing a row entirely due to a single missing response could result in more than a 5% truncation of the data set. Therefore, we decide if the missing rate percentage is within 0.5%, we remove the missing records rather than imputation. Otherwise, mean imputation was applied on the missing values.

2.3 Response Variable – Tenure

As seen in Figure 1, the distribution of tenure values is highly skewed.

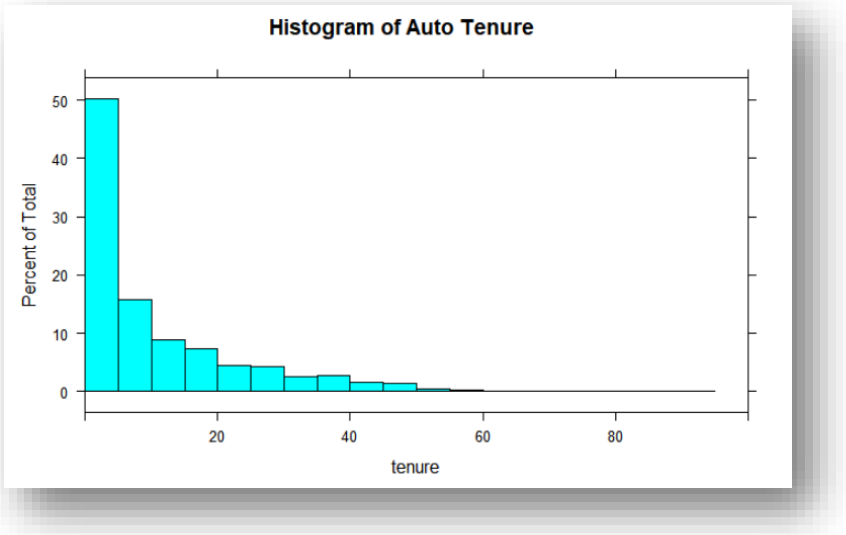

Figure 1: Histogram of Auto Tenure

For the interest of modeling and from a business perspective, we would like to cut the dataset into two sections: customers who stay with the insurer for more than 6 years vs. customers who stay with the insurer less than 6 years. Based on the bar chart below, we see there is almost an evenly distributed between the two groups.

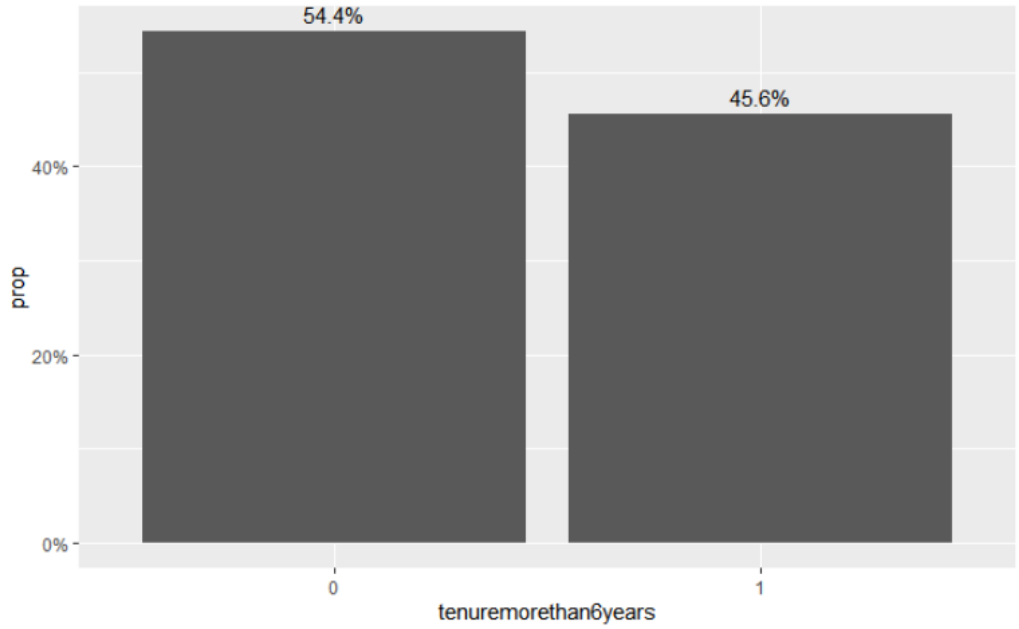

Figure 2: Bart Chart for Response Variable – Tenuremorethan6years

2.4 Variables Reduction – Correlation and Multicollinearity

I run the Pearson correlation to see the correlation relationship between the response variable tenure and the independent variables. By looking at the correlation, I remove the variables that are having the low correlations (<0.05) with the response variable tenure. This process reduced the variables from 100 to 42.

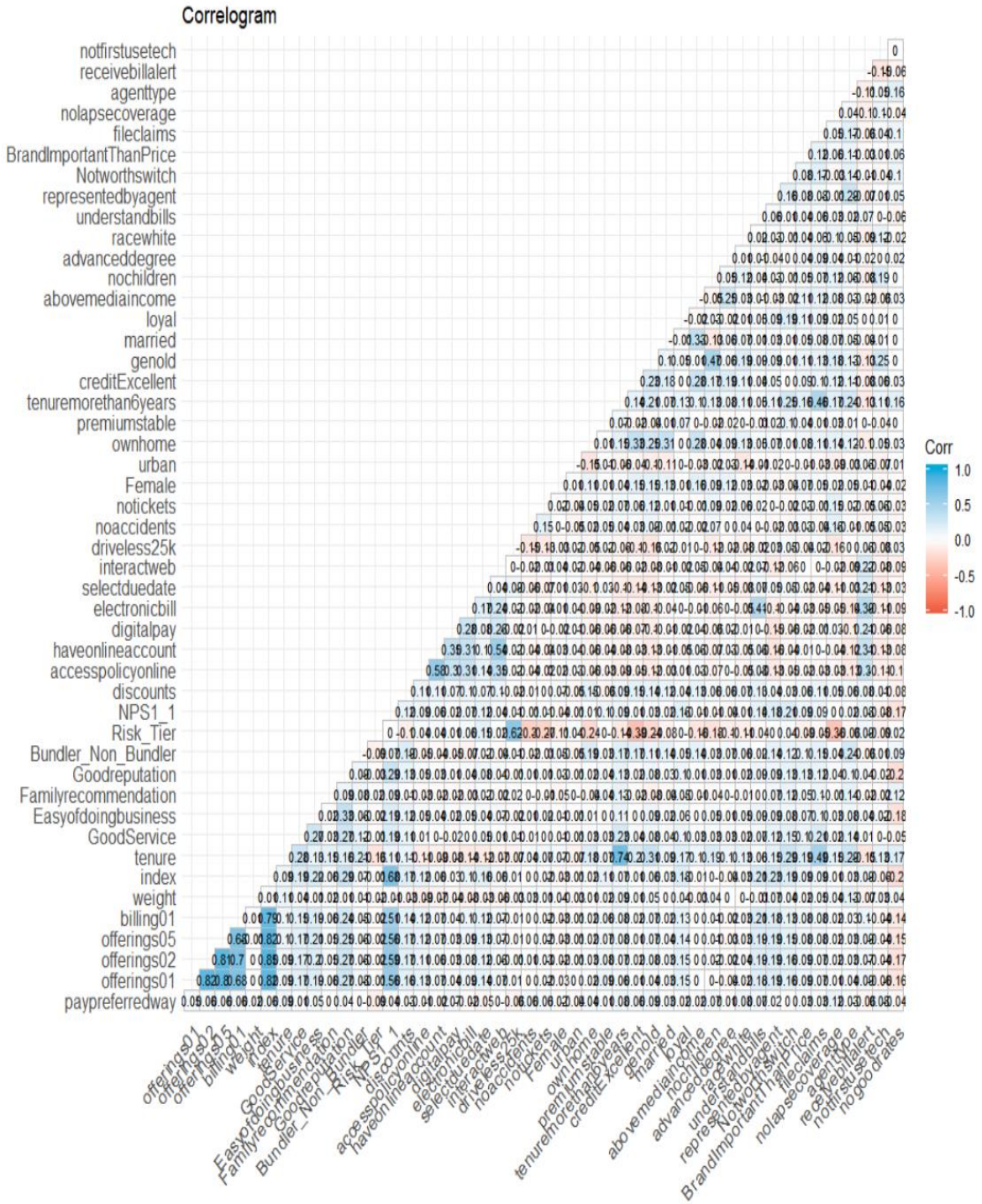

Figure 3: Variables Correlations Graph

Based on the correlation matrix above, some dependent variables are highly correlated with each other, which will create the multicollinearity issue. Having the redundant variables will present the issue of interpreting the models and overfitting the model. To reduce the multicollinearity, we have two methods [8]:

We can remove high correlated variables and keep one variable that represents similar features [8]. For example, index which measures the customers' overall satisfaction is highly correlated with offerings01 that measures the customers' satisfaction of policy offering. We keep the overall satisfaction Index since it is a broader measurement and more correlated to the response variable tenure that than of offerings01

We can also combine the highly correlated variables through PCA (Principal Component Analysis) [8]. The Principal component analysis can help reduce dimension and create a new independent variable based on the combination of existing 30+ variables [1].

2.5 Feature Engineering

2.5.1 Binarization

As part of the feature selection process, the binarization process is to create a new variable with only two values. The process is to remove redundant values and put the variables with similar values into the same group. Most of the raw variables were binarized through the exploratory analyses. For example, credit and generation variables were transformed into binary variables.

Credit level by tenure

As we would expect, customers tend to have longer tenure with better credit. By looking at their means, we see that customers with excellent credit have higher means than those who are not. On the other hand, "Good", "fair" and "poor" credit have similar means. Therefore, we

combined the “Good”, “Fair” and “Poor” into “not excellent credit” and have the “Excellent” on its own.

Generation by tenure

As we expect, as age goes, customers tend to be more stable and therefore stay longer with the company. As seen below, Gen X (Age), Gen Y (Age) and Gen Z (Age) have similar means. Therefore, we can group them into one bucket as young generation while having baby boomers and pre boomers roll into another bucket as older generation.

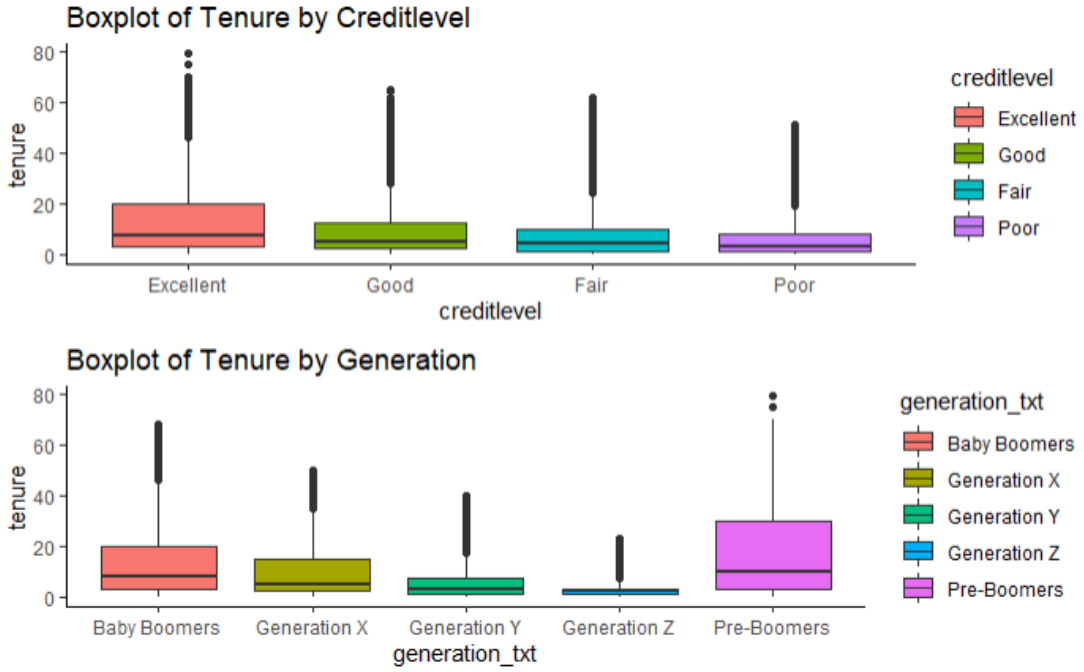

Figure 4: Boxplots of Credit Level and Generation

2.5.2 Seasonality

Based on the average tenure from 17-19, no significant trend is identified in average tenure more than 6 years rate. As we confirm there is no seasonality effect, year is dropped from the model.

Year	2017	2018	2019
Tenuremorethan6years	0.52	0.50	0.50

Table 1: Year Seasonality

2.5.3 Interaction Effect

We would like to see if there is any interaction effect within the independent variables. Our first thought would be looking at the relationship between above median income and premium stable. We assume more wealthy customers would be less sensitive with the price change. When above median income is 1 (”Yes”), tenure for customers who choose premium stable is higher than those who are not. While above median income is (“No”), tenure for customers is almost the same. Just like what we seen in “good service” and “easy of doing business”. We will use these two interaction effects for modeling later.

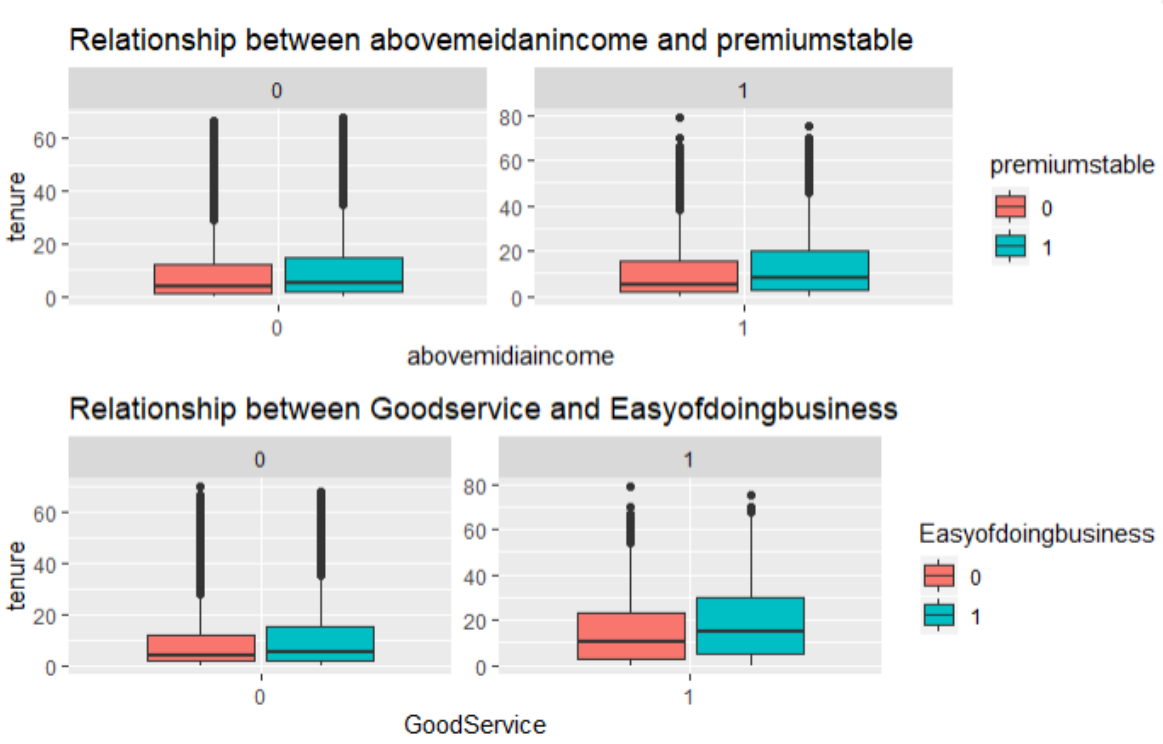

Figure 5: Interaction Effect

2.5.3 Data Descriptive

After the cleaning and feature engineering, the dataset can be grouped into three categories:

1. Demographic 2. Phycological 3. Ratings

Demographic:

GenOld: Generation (0 = Baby Boomer/Pre Bommers; Gen Y, Gen X and Gen Z)

Female: Gender is female (1=Yes; 0=No)

Driverlessthan25k: Miles less than 25k (1=Yes; 0=No)

Bundler_Non_Bundler: Policy is bundled (1=Yes; 0=No)

AdvancedDegree: People with an advanced degree (1=Yes; 0=No)

Abovemediaincome: Income is above median (1=Yes; 0=No)

Discounts: 2 or more discounts were offered (1=Yes; 0=No)

Racewhite: White (1= Yes; 0=No)

Ownhome: Have a home (1= Yes; 0=No)

Creditexcellent: Whether have an excellent credit or not (1= Yes; 0=No)

Nochildren: No children in the same household (1= Yes; 0=No)

Noaccidents: Have no more than 2 accidents (1= Yes; 0=No)

Married: Marriage status (1= Yes; 0=No)

Fileclaims: Status of file claims (0=file the claims; 1= file the claims less than 3 years; 2= file the claims more than 3 years)

Nolaspsecoverage: No coverage lapse (1= Yes; 0=No)

Understandbills: Completely understand bills (1= Yes; 0=No)

Prefferedway: Customers able to pay through preferred method (1= Yes; 0=No)

Premiumstable: Premium is stable (1= Yes; 0=No)

Risk: High/Median/Low risk (1= Low; 2= Median ;3= High)

Noshop: Did not shop (1 =Yes; 0 = No)

Noproblem: Did not experience problem (1 =Yes; 0 = No)

Agenttype: Types of agent (1= Independent agent; 0= no agent/exclusive agent)

Selectduedate: Select due date for bills (1 =Yes; 0 = No)

Recivebillalert: Receive bill alert (1 =Yes; 0 = No)

Accesspolicyonline: Access policy online (1 =Yes; 0 = No)

Phycological:

Loyal: Will definitely not change insurers (1= definitely not; 0= Probably not/Probably will/ definitely will) (1= Yes; 0=No)

OV9A: Important being represented by a local insurance agent (1=Very/Extremely; 0= Not at all/somewhat important) (1= Yes; 0=No)

CP6_R3: Good service experience in the past (1= Yes; 0=No)

CP6_R9: Good reputation (1= Yes; 0=No)

CP6_R6: Family/Friend Recommendation (1= Yes; 0=No)

CP6_R4: Convenience/Ease of doing business with (1= Yes; 0=No)

OV15d: Switching to another auto insurer is not worth the risk (1= Yes; 0=No)

OV15h: Brand name is more important to me than price (1= Yes; 0=No)

Ratings:

Index: Customer satisfaction (100-1000)

NPS: recommendation (1-10)

CHAPTER 3

3 Binary Classifications Modeling

In this section, we will focus on using logistic regression, decision trees, random forest and xgboost to predict the likelihood of customers staying with an insurer for more than 6 years.

I first split the dataset into the training (80% of the dataset) and testing dataset (20% of dataset). The response distribution is very similar in both training (53%) and testing dataset (47%). The base model is if our prediction were all wrong, there would be 50% accuracy. Thus, our model should have at least an accuracy of more than 50%.

For logistic regression, I used the backward logistic regression as a baseline model. To improve the accuracy of the baseline model, I performed the following method: K-fold cross validation, adding the interaction effect terms and PCA. The best accuracy of the logistic regression was compared to the accuracy of the decision trees, random forest and xgboost.

3.1 Data Standardization

Before we fit the data into a model, we need to ensure data is standardized. Since most of our exploratory variables are already binarized, the only two numerical variables that should be standardized are index and NPS. While logistic regression and ensemble trees algorithms will not require standardization, the performance of other machine learning algorithms may be negatively impacted by non-scaled data [12].

3.2 Logistic Regression

Logistic regression is a basic algorithm to deal with the classification problem. For the tenure problem, the logistic regression models the probability of tenure for more than 6 years. The outcome probability is between the value of 0 and 1. In default, if the probability of tenure more than 6 years is > 0.5, then we classify this customer to be staying with an insurer for more than 6 years.

3.2.1 Backward Selection

Since we have 30+ variables in the dataset, we should consider removing the unnecessary variables. By applying the backward selection, the method starts with all 30+ variables and then eliminates the least important variables [11]. After looking at the summary of the final chosen model, every variable is significant (p-value is less than 0.1).

The accuracy of training: 0.77 The accuracy of test:0.77

This means we can use the logistic regression model to correctly predict 77% of the customers staying with the same insurer for more than 6 years on a different dataset.

```
Start:  AIC=3248166
tenuremorethan6years ~ index + GoodService + Easyofdoingbusiness +
    Familyrecommendation + Goodreputation + Bundler_Non_Bundler +
    discounts + accesspolicyonline + electronicbill + selectduedate +
    driveless25k + noaccidents + notickets + Female + ownhome +
    premiumstable + creditExcellent + genold + loyal + abovemediaincome +
    nochildren + advanceddegree + racewhite + understandbills +
    representedbyagent + Notworthswitch + BrandImportantThanPrice +
    fileclaims + nolapsecoverage + agenttype + receivebillalert +
    +paypreferredway + notfirstusetech + nogoodrates + NPS1_1 +
    fileotherclaims + risk + noshop + noproblem

                          Df Deviance     AIC
<none>                         3248082 3248166
- notickets                1  3248105 3248187
- accesspolicyonline       1  3248140 3248222
- discounts                1  3248205 3248287
- representedbyagent       1  3248344 3248426
- Bundler_Non_Bundler      1  3248431 3248513
- Female                   1  3248594 3248676
- advanceddegree           1  3248820 3248902
- noaccidents              1  3248965 3249047
- abovemediaincome         1  3249020 3249102
- creditExcellent          1  3249027 3249109
- racewhite                1  3249303 3249385
- ownhome                  1  3249414 3249496
- notfirstusetech          1  3249442 3249524
- nochildren               1  3249467 3249549
- driveless25k             1  3249484 3249566
- paypreferredway          1  3249673 3249755
- receivebillalert         1  3249843 3249925
- noproblem                1  3250336 3250418
- NPS1_1                   1  3250421 3250503
- Easyofdoingbusiness      1  3251508 3251590
- understandbills          1  3252165 3252247
- index                    1  3252605 3252687
- Goodreputation           1  3252666 3252748
- electronicbill           1  3253163 3253245
- premiumstable            1  3253609 3253691
- loyal                    1  3254104 3254186
- genold                   1  3254261 3254343
- BrandImportantThanPrice  1  3254270 3254352
- selectduedate            1  3254876 3254958
- risk                     2  3257563 3257643
- Familyrecommendation     1  3258142 3258224
- agenttype                1  3264463 3264545
- fileotherclaims          1  3273636 3273718
- nogoodrates              1  3284702 3284784
- GoodService              1  3284749 3284831
- nolapsecoverage          1  3295201 3295283
- Notworthswitch           1  3301489 3301571
- noshop                   1  3337472 3337554
- fileclaims               2  3416844 3416924
```

Figure 5: Stepwise Logistic Regression

3.2.2 Compare Residual Deviance

Anova analysis was run on the logistic regression to detect the residual changes for each variable. As seen in the Figure 6, adding *fileclaims*, *Goodservice* and *notworthswitch* are

significantly reduces the residual deviance. While the variable *married* barely reduces the deviance. To make the model simpler, we removed the variable *married.*

By removing the married, the updated AIC is slightly higher than the base model. The model was chosen over the initial backward model given the accuracy for both training and testing is similar, but this model has less exploratory variables which simplifies the model.

The accuracy of training: 0.765 The accuracy of the test:0.767

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			77412	4853537		
index	1	18524	77411	4835013	< 2.2e-16	***
GoodService	1	171422	77410	4663591	< 2.2e-16	***
Easyofdoingbusiness	1	9973	77409	4653618	< 2.2e-16	***
Familyrecommendation	1	57249	77408	4596369	< 2.2e-16	***
Goodreputation	1	10317	77407	4586053	< 2.2e-16	***
Bundler_Non_Bundler	1	66713	77406	4519339	< 2.2e-16	***
discounts	1	2506	77405	4516834	< 2.2e-16	***
accesspolicyonline	1	30534	77404	4486299	< 2.2e-16	***
electronicbill	1	36769	77403	4449530	< 2.2e-16	***
selectduedate	1	44480	77402	4405050	< 2.2e-16	***
driveless25k	1	10618	77401	4394432	< 2.2e-16	***
noaccidents	1	3603	77400	4390830	< 2.2e-16	***
notickets	1	8722	77399	4382107	< 2.2e-16	***
Female	1	4942	77398	4377165	< 2.2e-16	***
ownhome	1	29849	77397	4347317	< 2.2e-16	***
premiumstable	1	14083	77396	4333234	< 2.2e-16	***
creditExcellent	1	11626	77395	4321608	< 2.2e-16	***
genold	1	66281	77394	4255327	< 2.2e-16	***
loyal	1	30764	77393	4224563	< 2.2e-16	***
abovemediaincome	1	8685	77392	4215878	< 2.2e-16	***
nochildren	1	5163	77391	4210715	< 2.2e-16	***
advanceddegree	1	4281	77390	4206433	< 2.2e-16	***
racewhite	1	6427	77389	4200006	< 2.2e-16	***
understandbills	1	9655	77388	4190352	< 2.2e-16	***
representedbyagent	1	3076	77387	4187275	< 2.2e-16	***
Notworthswitch	1	137622	77386	4049653	< 2.2e-16	***
BrandImportantThanPrice	1	19213	77385	4030441	< 2.2e-16	***
fileclaims	2	501021	77383	3529420	< 2.2e-16	***
nolapsecoverage	1	56055	77382	3473365	< 2.2e-16	***
agenttype	1	30657	77381	3442708	< 2.2e-16	***
receivebillalert	1	5079	77380	3437629	< 2.2e-16	***
paypreferredway	1	2726	77379	3434904	< 2.2e-16	***
notfirstusetech	1	5766	77378	3429138	< 2.2e-16	***
nogoodrates	1	44769	77377	3384368	< 2.2e-16	***
NPS1_1	1	3097	77376	3381271	< 2.2e-16	***
fileotherclaims	1	30070	77375	3351201	< 2.2e-16	***
risk	2	11040	77373	3340161	< 2.2e-16	***
noshop	1	89824	77372	3250336	< 2.2e-16	***
noproblem	1	2255	77371	3248082	< 2.2e-16	***
married	1	590	77370	3247491	< 2.2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure 6: Residual Deviance

3.2.3 Check Multi-collinearity

From the correlation table, we see there is some correlation between predictors. While the correlations are not high between the predictors, we still want to ensure there is no multicollinearity issue within the model. Variance inflation factor (VIF) is used to assess multi-collinearity [5]. In general, if VIF is bigger than 4, then we should consider removing the variable from the model [5]. VIF for *drivelessthan25k* and *risk* is bigger than 4, we decide to remove the variable "driveless25k". This is because *driveless25k* has smaller residual deviance and is less important for contributing to the logistic regression model. As seen in figure 7 and 8 below with *driveless25k* removed, all VIF factors are above 4.

	GVIF	Df	GVIF^(1/(2*Df))
index	2.133016	1	1.460485
GoodService	1.136502	1	1.066068
Easyofdoingbusiness	1.191140	1	1.091394
Familyrecommendation	1.067609	1	1.033252
Goodreputation	1.272264	1	1.127947
Bundler_Non_Bundler	1.156698	1	1.075499
discounts	1.167523	1	1.080520
accesspolicyonline	1.215889	1	1.102674
electronicbill	1.522976	1	1.234089
selectduedate	1.140850	1	1.068106
driveless25k	4.266899	1	2.065647
noaccidents	1.087308	1	1.042740
notickets	1.402400	1	1.184230
Female	1.088288	1	1.043211
ownhome	1.257620	1	1.121437
premiumstable	1.056727	1	1.027972
creditExcellent	1.269769	1	1.126840
genold	1.524593	1	1.234744
loyal	1.067135	1	1.033022
abovemediaincome	1.246096	1	1.116287
nochildren	1.330579	1	1.153507
advanceddegree	1.095185	1	1.046511
racewhite	1.061959	1	1.030514
understandbills	1.306667	1	1.143095
representedbyagent	1.184107	1	1.088167
Notworthswitch	1.127341	1	1.061763
BrandImportantThanPrice	1.043599	1	1.021567
fileclaims	1.944395	2	1.180854
nolapsecoverage	1.088986	1	1.043545
agenttype	1.157164	1	1.075715
receivebillalert	1.276846	1	1.129976
paypreferredway	1.033610	1	1.016666
notfirstusetech	1.141228	1	1.068283
nogoodrates	1.184731	1	1.088453
NPS1_1	1.934012	1	1.390688
fileotherclaims	1.115298	1	1.056077
risk	8.679678	2	1.716429
noshop	1.099691	1	1.048661
noproblem	1.075052	1	1.036847

Figure 7: VIF for Logistic Regression with All Variables

	GVIF	Df	GVIF^(1/(2*Df))
index	2.131820	1	1.460075
GoodService	1.136321	1	1.065984
Easyofdoingbusiness	1.191161	1	1.091403
Familyrecommendation	1.066886	1	1.032902
Goodreputation	1.271876	1	1.127775
Bundler_Non_Bundler	1.156417	1	1.075368
discounts	1.167454	1	1.080488
accesspolicyonline	1.215934	1	1.102694
electronicbill	1.522436	1	1.233870
selectduedate	1.140901	1	1.068130
noaccidents	1.082142	1	1.040260
notickets	1.180504	1	1.086510
Female	1.088272	1	1.043203
ownhome	1.256515	1	1.120944
premiumstable	1.056699	1	1.027959
creditExcellent	1.264997	1	1.124721
genold	1.523509	1	1.234305
loyal	1.067067	1	1.032989
abovemediaincome	1.245975	1	1.116232
nochildren	1.326936	1	1.151927
advanceddegree	1.095064	1	1.046453
racewhite	1.061871	1	1.030471
understandbills	1.306697	1	1.143108
representedbyagent	1.183287	1	1.087790
Notworthswitch	1.127142	1	1.061669
BrandImportantThanPrice	1.043418	1	1.021479
fileclaims	1.920792	2	1.177254
nolapsecoverage	1.079010	1	1.038754
agenttype	1.156935	1	1.075609
receivebillalert	1.276766	1	1.129941
paypreferredway	1.033430	1	1.016578
notfirstusetech	1.141327	1	1.068329
nogoodrates	1.184566	1	1.088378
NPS1_1	1.933126	1	1.390369
fileotherclaims	1.112735	1	1.054862
risk	2.213477	2	1.219744
noshop	1.099552	1	1.048595
noproblem	1.074676	1	1.036666

Figure 8: VIF for Logistic Regression after Removing “driveless25k”

3.3 Logistic model performance

We will focus on assessing the logistic regression model performance and implementing three approaches to improve the logistic regression model accuracy.

3.3.1 Assess Model Performance

We assessed the model performance from two perspectives. The accuracy of the training (0.765) and testing dataset (0.767) were similar, which indicates that the model is not overfitting.

Moreover, we looked at the ROC and AUC (area under the curve) performance of logistic regression.

The confusion metrics of this problem can be defined as below:

	Actual "Tenure is more than 6 years"	Actual "Tenure is less than 6 years"
Predicted "Tenure is more than 6 years"	True Positives	False Positives
Predicted "Tenure is less than 6 years"	False Negatives	True Negatives

Table 2: Confusion Metrics

$$Ture\ Positive\ Rate = \frac{True\ postives}{True\ postives + False\ nagatives}$$

$$False\ Positive\ Rate = \frac{False\ postives}{False\ postives + True\ nagatives}$$

When the threshold set at 0.5, the accuracy of the testing dataset is 0.767 and AUC=0.763. We use AUC to help decide which method is better. Normally the large AUC implies the better model. Below is a graph of the ROC with a threshold at 0.5. The Y axis shows the true positive rate, which illustrates the proportion of more than 6 years that were correctly classified. The X axis shows the false positive rate, which illustrates the proportion of less than 6 tenure that were incorrectly classified.

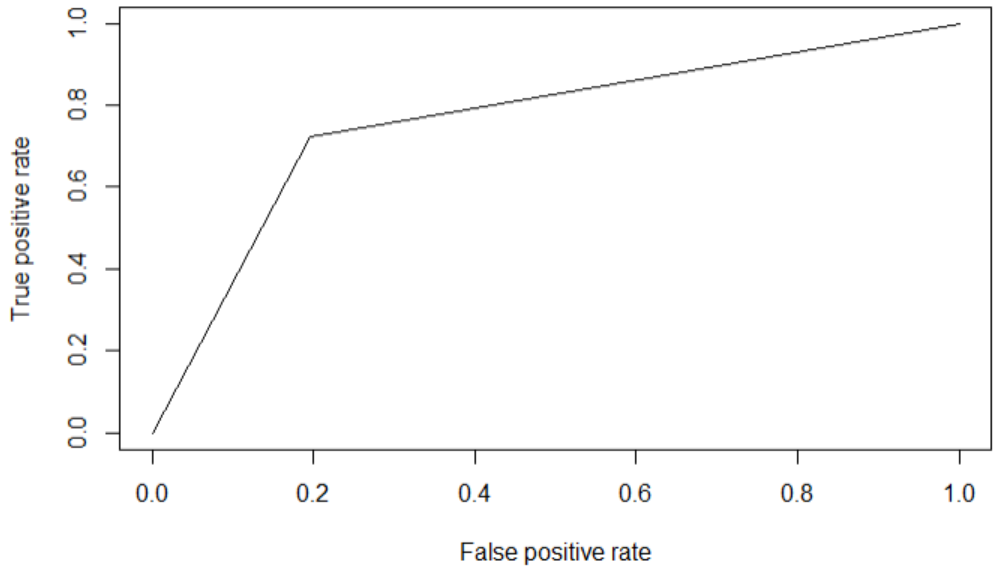

Figure 8: ROC Graph

If we change the threshold to 0.55 and 0.6 respectively, the accuracy of the testing dataset is slightly improved. However, AUC curve does not seem to get better. As such, we still maintain the 0.5 threshold as we initially set.

Threshold	Accuracy of testing dataset	AUC
0.5	0.765	0.763
0.55	0.768	0.760
0.6	0.769	0.757

Table 3: Summary of AUC and Accuracy at Different Level of Threshold

3.3.2 Methods of Improving Model Accuracy

We tested the following methods to improve the base model logistic regression accuracy.

1. Adding the interaction effect
2. K-folder cross validation
3. Conduct the PCA analyses

3.3.2.1 Adding Interaction Effect

We added the interaction terms that we identified in the exploratory analyses previously. We ran the model on the training and test dataset. While two interaction terms are significant in the model, the accuracy was only slightly improved with the additional two interaction terms.

The accuracy of the training: 0.7663. The accuracy of the testing: 0.7677

3.3.2.2 K-fold Validation

The K-fold cross validation is to divide the data into K folds. Of those K-folds, one fold will be used for testing dataset while the rest of K-1 folds will be used to train the data. If we are doing 500 K-fold cross validation, model would will repeat the process 500 times [10]. To understand how each iteration performs, we plot the histogram of 500-fold cross validation based on the final reduced model. As seen in the boxplot below, the accuracy scores of the model are all above 0.7 other than the two outliers. The mean of accuracy is: 0.762

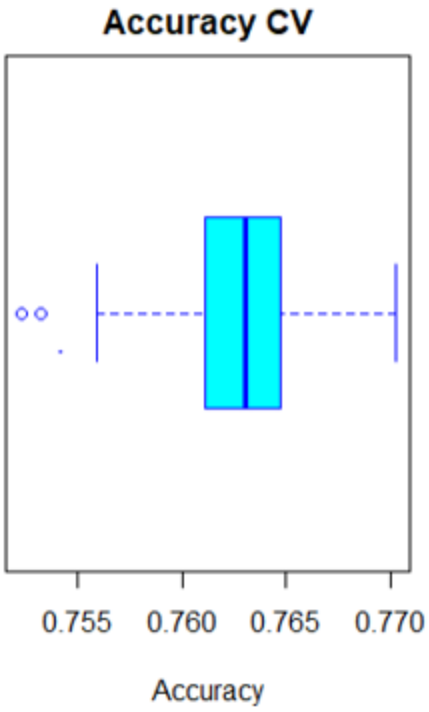

Figure 9: K-fold Cross Validation Accuracy Scores

3.3.2.3 PCA (Principal Component Analyses)

We worked with 30+ variables. Some of them are highly correlated with each other. PCA will help to reduce the multicollinearity issue within variables and avoid overfitting the model. By applying PCA, we will create new independent variables that are the linear combination of those 30+ exploratory variables [1]. Instead of dropping insignificant variables as we did in backward elimination, we will still retain all the variables to prevent losing any important information through PCA [1].

Below is the result of eigen values – we choose the principal components with eigen values that are bigger than 1. Thus, we include the first 10 PCs in our dataset. The first 10 PCs explain 47% of the variance.

```
 [1] 3.3308652 2.9879538 2.0294058 1.9341302 1.4051793 1.3463890 1.2291050 1.1423656
 [9] 1.0942480 1.0176305 0.9998239 0.9497413 0.9404312 0.9119335 0.9029834 0.8868546
[17] 0.8765436 0.8560351 0.8328347 0.8252526 0.8125031 0.7778560 0.7686587 0.7430175
[25] 0.7307103 0.7035429 0.6943068 0.6832390 0.6609699 0.6510247 0.6422908 0.6131267
[33] 0.5806292 0.4504517 0.4263678 0.3070131 0.2545854
```

eig <dbl>	variance <dbl>	cumvariance <dbl>
3.3308652	9.0023383	9.002338
2.9879538	8.0755508	17.077889
2.0294058	5.4848806	22.562770
1.9341302	5.2273790	27.790149
1.4051793	3.7977819	31.587931
1.3463890	3.6388892	35.226820
1.2291050	3.3219054	38.548725
1.1423656	3.0874745	41.636200
1.0942480	2.9574270	44.593627
1.0176305	2.7503528	47.343979

Figure 10: Eigen Values for PCA

By including the first 10 PCs in the dataset and running the logistic regression model to predict the response variable tenuremorethan6years, all 10 PCs variables are significant.

The accuracy of training: 0.7465 The accuracy of testing: 0.7502

```
Call:
glm(formula = tenuremorethan6years ~ PCA$x[, 1] + PCA$x[, 2] +
    PCA$x[, 3] + PCA$x[, 4] + PCA$x[, 5] + PCA$x[, 6] + PCA$x[,
    7] + PCA$x[, 8] + PCA$x[, 9] + PCA$x[, 10], family = binomial
    data = train, weights = train$weight)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-44.280   -4.771   -1.591    4.417   40.087

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.0392622  0.0013042  -30.11   <2e-16 ***
PCA$x[, 1]   0.5368981  0.0008260  650.03   <2e-16 ***
PCA$x[, 2]  -0.2480932  0.0007692 -322.52   <2e-16 ***
PCA$x[, 3]  -0.3526649  0.0009513 -370.71   <2e-16 ***
PCA$x[, 4]  -0.3352985  0.0009455 -354.63   <2e-16 ***
PCA$x[, 5]  -0.0233657  0.0010896  -21.44   <2e-16 ***
PCA$x[, 6]   0.4226139  0.0011452  369.04   <2e-16 ***
PCA$x[, 7]  -0.3756643  0.0012084 -310.89   <2e-16 ***
PCA$x[, 8]   0.0848992  0.0012130   69.99   <2e-16 ***
PCA$x[, 9]   0.4751236  0.0012955  366.75   <2e-16 ***
PCA$x[, 10]  0.0877669  0.0012763   68.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4853537  on 77412  degrees of freedom
Residual deviance: 3575906  on 77402  degrees of freedom
AIC: 3575928

Number of Fisher Scoring iterations: 7
```

Figure 11: PCA R Outputs

3.5 Decision Tree

Decision Tree is very easy to interpret and visualize compared to other machine learning methods. There are four important parameters that we should specify when we built the initial tress. Those are Minsplit. Minbucket, Cp and Maxdepth.

While we can build a very complicated tree to gain a better accuracy, we should be careful not overfitting the model. I went extreme and started with the tree with a small minsplit =20, cp=0, minbucket=20 and maxdepth=30. The accuracy for the training dataset is: 0.85. The accuracy for

the testing dataset is: 0.69. There is a big difference in accuracy scores between the training and testing dataset. I concluded that the model is overfitting. To resolve the overfitting issue, I pruned the tree back by looking at the smallest CP.

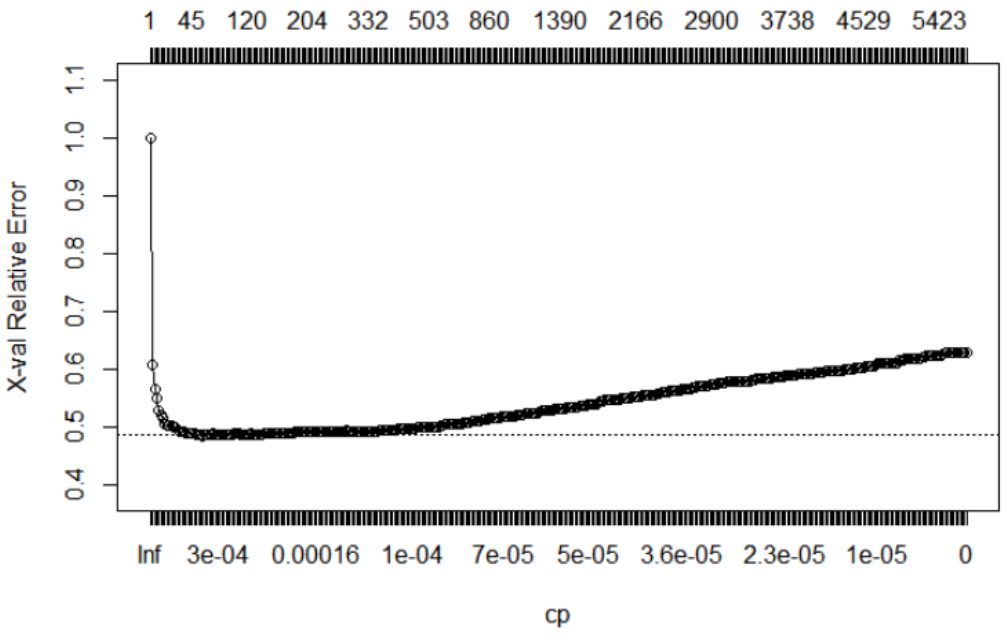

Figure 12: Complexity Parameter for Decision Tree

The minimal cp is 0.0003061631. Replace cp=0 with the minimal cp.

The updated accuracy of the training dataset: 0.765 The updated accuracy of the testing dataset: 0.760

3.6 Random forest

Random forest is built based on many decision trees [4]. To gain better accuracy, random forest averages all outcomes of the trees to reduce the variance [4]. In comparison to a single decision tree, random forest is much less interpretable given thousands of trees are produced.

We started with the base model with the default parameters: number of trees=500, number of variables =6. Below is the plot of our base random forest model. As we can see, after ntree=400, the error is barely changed. Thus, having more trees not improve the accuracy

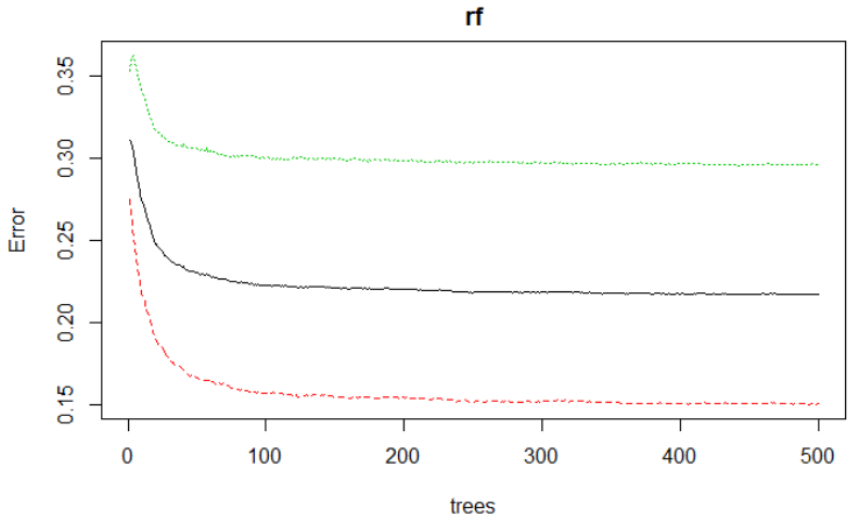

Figure 13: Out of Bag Error for Number of Trees

Next, we looked at the optimum number of features we can choose. The error is at the lowest when mtry=6.

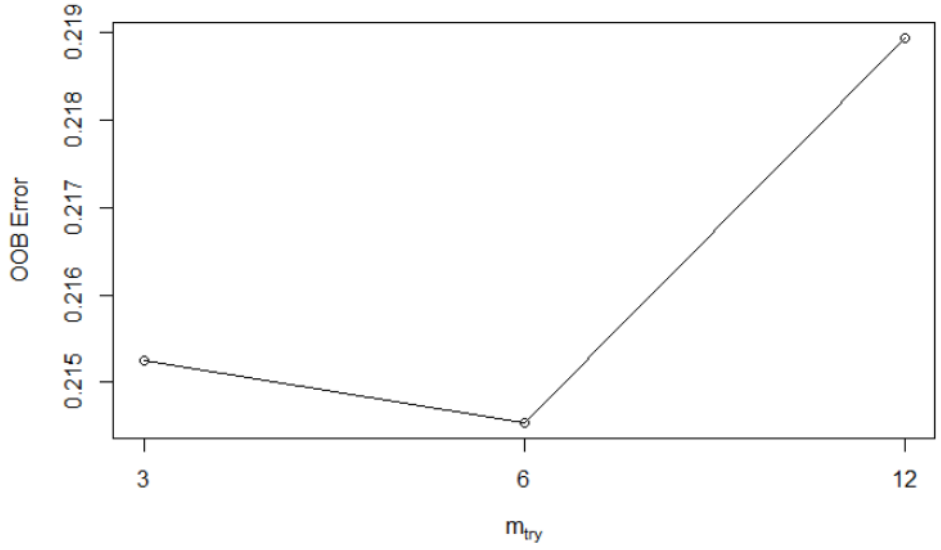

Figure 14: Out of Bag Error for Number of Variables

The accuracy for the training is: 0.998 The accuracy for the testing is: 0.79. The model is overfitting.

3.7 xgboost

Xgboost is one of the ensemble learning algorithms. I used the xgboost package in R to implement the xgboost model.

Through the cross-validation, the prediction power reaches the best when the rounds =226. We set the rounds=226, the max_depth=10 and eta =0.3. The accuracy of training dataset= 0.84 The accuracy of testing dataset= 0.777

Below is the xgboost importance features. The top 20 importance features are plotted. It is interesting to show no problem is the most important feature.

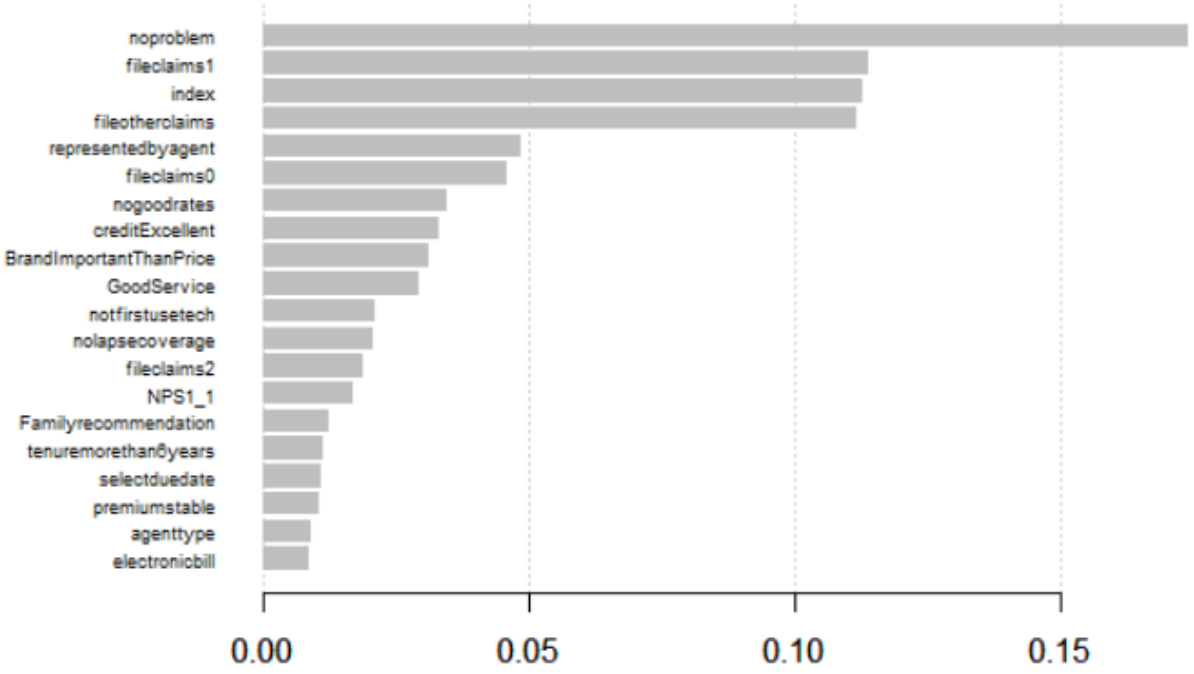

Figure 15: Feature Importance for Xgboost

3.8 Conclusion

As seen in the table below, random forest is the best model which slightly improve the accuracy by 0.01 compared to the logistic regression.

Models	Accuracy of training	Accuracy of testing
Backward logistic regression	0.765	0.767
Adding interaction – Backward logistic regression	0.766	0.777
K-fold cross validation	0.76	0.762
Principal Component Analyses	0.75	0.75
Decision Trees	0.765	0.76
Random forest	0.99	0.79
Xgboost	0.84	0.78

Table 4: Summary of All Methods Training and Testing Accuracy

This means we can use the random forest model to correctly predict 79% of the customers staying with the same insurer for more than 6 years on a different dataset.

CHAPTER 4

4 Model Interpretation and Data Analysis

In the section, I focus on understanding the key factors for customers staying with the same insurer for more than 6 years based on the logistic regression and decision trees modeling results from the previous chapter.

By looking at the odds ratio of the multinomial logistic regression, this will help us understand if there are any difference in customers' behaviors between the tenure [0-3] years vs. [4-6] years vs. [6 +] years. Moreover, any commonality in generations in tenure will be further investigated.

4.1 Models Interpretation - Logistic Regression

In this section, we focus on interpreting the binary and multinomial logistic regression modeling results.

4.1.1 Binary Logistic Regression

Based on the previous logistic regression, below is the odds ratio of the binary logistic regression. The top 5 odds ratio are *fileclaims2*, *nolapsecoverage*, *fileclaims1*, *have not shopped* and *file other claims*. Here is the model interpretation:

- The odds of having tenure >=6 years is 10 times for customers who have filed the claims more than 3 years than those who have not filed the claims.
- The odds of having tenure >=6 years is 7.5 times for customers who have no lapse coverage than those who had the lapse coverage.

- The odds of having tenure >=6 years is 2.4 times for customers who have filed the claims less than 3 years than those who have not filed the claims.
- The odds of having tenure >=6 years is 2.4 times for customers who have not shopped in the past 12 months than those who have shopped.
- The odds of having tenure >=6 years is 2.3 times for customers who feel it is not worth switch to insurers than those who are.
- The odds of having tenure >=6 years is 2 times for customers have other claims besides auto claims than those who are not.

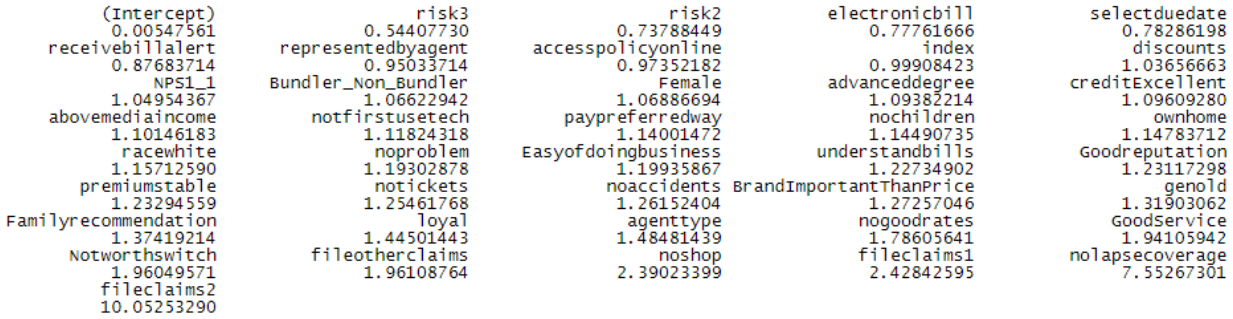

Figure 16: Odds Ratio for Binary Logistic Regression

Below is the variable importance table. The most important 5 variables that contribute to the models are the varaibles *fileclaims more than 3 years*, *file claims less than 3 years*, *not worth to switch*, *no good rates* and *no lapse coverage*.

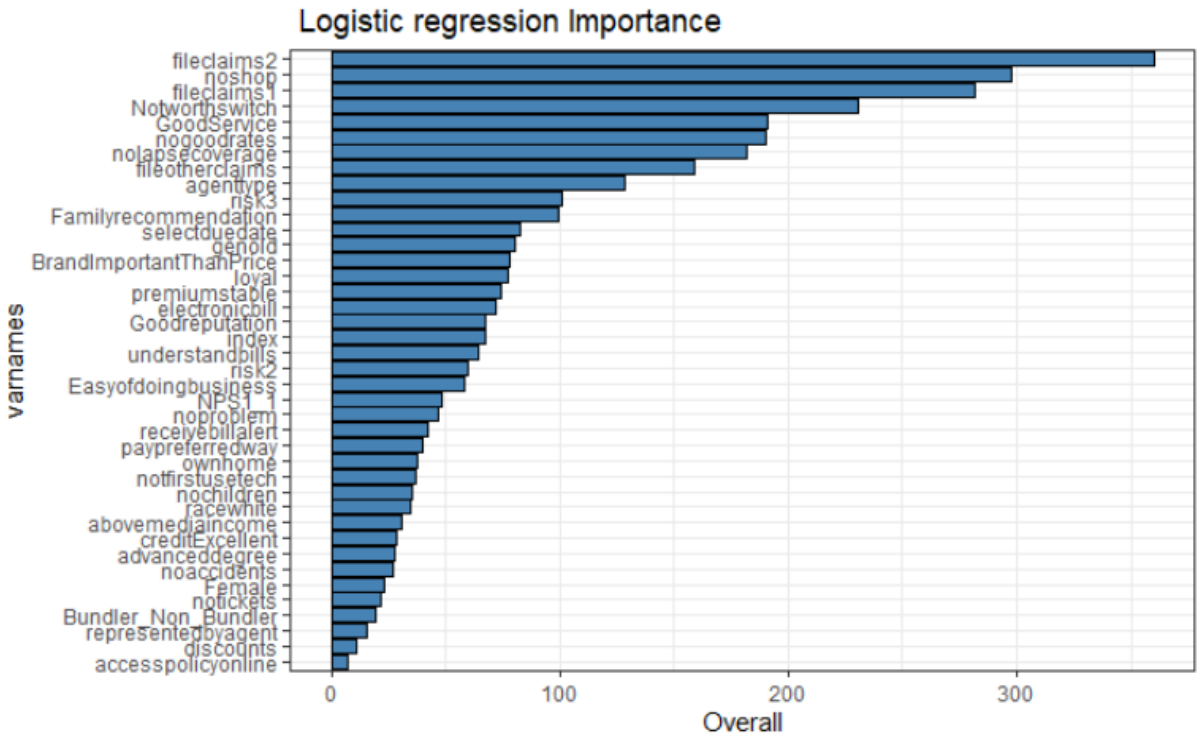

Figure 20 : Variable Importance for Logistic Regression

4.1.2 Multi-class Logistic Regression

The multinomial logistic regression was conducted to determine what is the likelihood for customers staying with the same insurer in [0-3] years vs. [3-6] years vs. [7 years+]. Below is a summary of the odds ratio for multinomial logistic regression. To set [4-6] years as a baseline, below is the odds ratio table. The interpretation of the odds ratio for good service is as below.

Keeping all other variables consistent, when good service increases one unit, it is 64% more likely tenure is more than 6 years vs. tenure is 4 to 6 years. Keeping all other variables consistent, when good service increases one unit, it is 20% less likely tenure is 0-3 years vs. tenure is 4 - 6 years.

Odds ratios for file claims, no lapse coverage, not worth to switch, good service, no tickets, premium stable and agent type are significantly lower in [0-3] and higher in [6 years+] meaning that those are key factors for customers staying with the insurer longer.

In looking at the odds ratio between [tenure more than 6 years] vs. [tenure between 3-6 years], having independent agent (odds is 54% higher), being loyal (odds is 49% higher) and getting family recommendation (odds is 38% higher) are also the important factors to drive customers from moving from tenure 4 to 6 years to tenure more than 6 years. On the other hand, having electronic bills (odds is 22% less) and selecting due date (odds is 16% less) will badly impact on customers with longer tenure.

By comparing the odds ratio between [tenure between 0 and 2 years] vs. [tenure between 3-6 years], having good reputation (odds is 11% less), having discounts (odds is 13% less), having no accidents (odds is 18% less), better understanding bills (odds is 17% less) and paying through the preferred way (odds is 15% less) are the important drivers to differentiate tenure between [4-6 years] vs. [0-3 years].

When customers do not feel have good rates (nogoodrates), they are less likely to stay in [0-3 years]. Therefore, this may imply that having good rates is important to attract customers in the beginning. Although the variable *not having good rates* is less important in keeping customers staying longer for more than 6 years.

```
  (Intercept)     index GoodService Easyofdoingbusiness Familyrecommendation Goodreputation
1 26.04704744 1.0007988    0.803559           0.9352347            0.9431783      0.8874549
3  0.04443905 0.9995538    1.646064           1.1261181            1.3856392      1.1446842
  Bundler_Non_Bundler discounts accesspolicyonline electronicbill selectduedate
1           0.9810372 0.8718514          1.0345926      1.1009468      1.183904
3           1.1542583 0.9980597          0.9719051      0.7886618      0.842805
  driveless25k noaccidents notickets    Female  ownhome premiumstable creditExcellent
1    1.0279176   0.8234056 0.7542935 0.9326677 1.004985     0.7771566        1.004654
3    0.8854066   1.1866351 1.3704589 0.9943546 1.167546     1.1424772        1.128708
     genold     loyal abovemediaincome nochildren advanceddegree racewhite understandbills
1 0.9603313 0.9742912         1.012343  0.9236088       1.037491 0.8878621       0.8377716
3 1.4078441 1.4986046         1.136631  1.1454628       1.129664 1.1861143       1.0835537
  representedbyagent Notworthswitch BrandImportantThanPrice fileclaims1 fileclaims2
1          1.0981426      0.7185592               0.9559845   0.4886612   0.1982555
3          0.9872662      1.7767318               1.2319435   1.7307621   6.3919949
  nolapsecoverage agenttype receivebillalert paypreferredway notfirstusetech nogoodrates
1       0.2822438 0.9727485        1.0641800       0.8536214       0.9093903   0.7871019
3       4.2124110 1.5420467        0.8486762       1.0918413       1.1950172   1.6194130
     NPS1_1
1 0.9547554
3 1.0265188
```

Figure 17: Odds Ratios for Multinominal Logistic Regression

4.2 Models Interpretation - Decision trees

Below is the initial decision tree when I set the parameters minsplit=3000, cp=0 , maxdepth=5 and minbucket=3000.

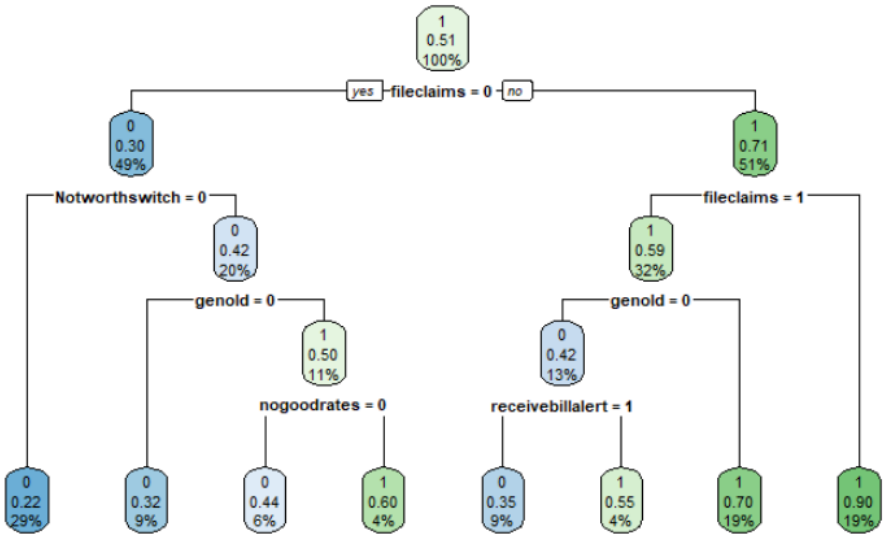

Figure 18: Decision Trees

The model shows that if customers have not filed a claim and choose no for the question "Not worth to switch insurer", there is a 22% chance they would stay with the same company more

than 6 years. If pre/baby boomers' customers have not filed a claim and choose No for the question "Not worth to switch insurer", there is a 32% chance they would stay with the same company for more than 6 years. The third lowest probability is the group who are Gen x/Gen y/Gen z, submit claims for more than 3 years and receive the bill alert, there is a 35% chance their tenure would be more than 6 years. This is implying that younger generations are sensitive to be informed with billings.

On the other hand, for those who have file the claims more than 3 years, there are a 90% chance that they are likely to stay with the company for more than 6 years. Then there is a 70% chance that customers would stay with the insurer for more than 6 years if pre/baby boomers have filed the claims less than 3 years.

In other words, customers who filed the claims for more than 3 years have the highest probability of staying with the insurer for more than 6 years. This finding is consistent with the findings in logistic regression.

Variable importance shows the ordering of variables based on their contribution to the model.

```
Variable importance
         fileclaims      Notworthswitch              genold        GoodService          agenttype
                 51                  10                   8                   6                   5
Bundler_Non_Bundler    abovemediaincome          nochildren            ownhome          Risk_Tier
                  5                   5                   3                   1                   1
    notfirstusetech    receivebillalert     creditExcellent        nogoodrates
                  1                   1                   1                   1
```

Figure 19: Variables Importance for Decision Tree

Similar to the results we see in that of logistic regression, the variable *fileclaims* is the most important factor then followed by the variable *norworthswitch*.

4.3 Data Analysis - Generational Difference

Since generation (Genold) is the top-level factor for decision trees, it would be interesting to see if there is any generational difference when customers decide staying with an insurer for more than 6 years. Thus, I took the generation variable out from the model and then fit the data into a decision model separately for the younger generation (Gen X/Gen y/Gen Z) and older generation (Pre/Baby boomers).

For both generations, we see the probability with tenure more than 6 years is 10% higher for those who feel the company with a good reputation vs. not. If we interpret "worth to switch" are those who are shopping around, then having a good reputation is vital to keep the older generation

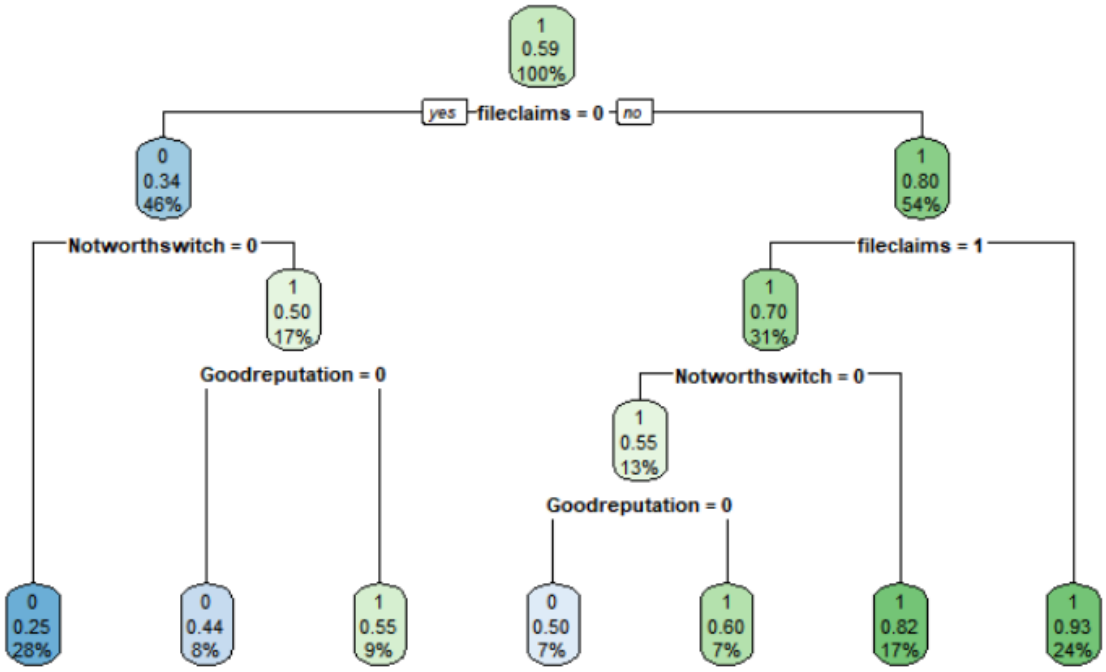

Figure 20: Decision Trees for Older Generation

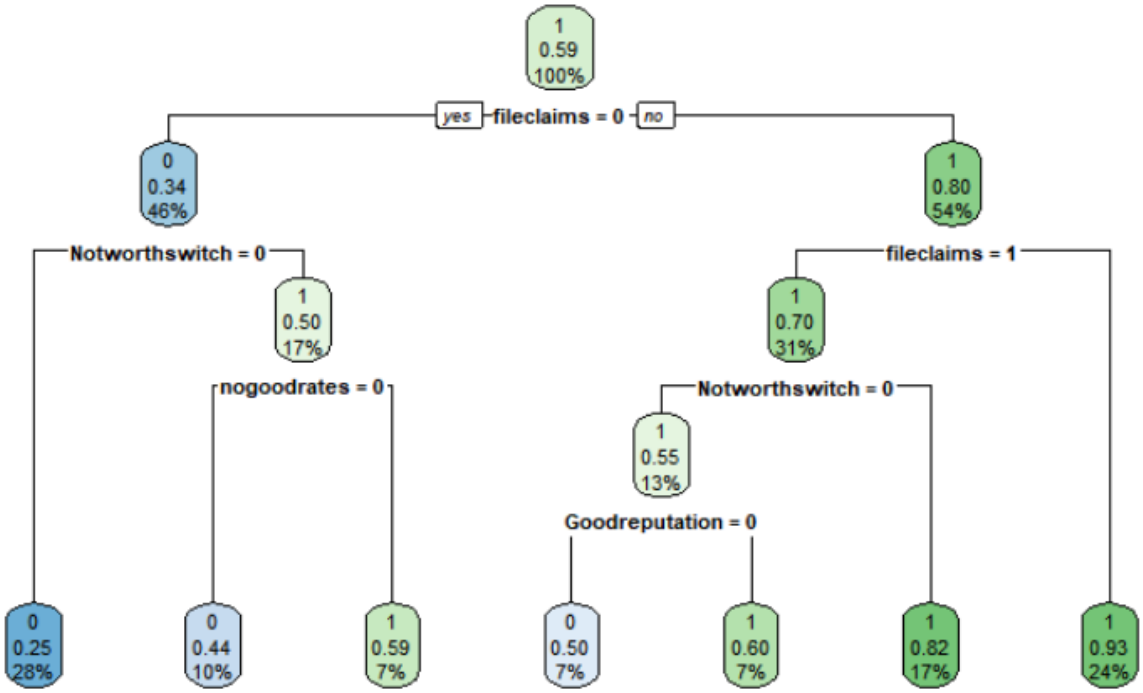

Figure 25: Decision Trees for Younger Generation

CHAPTER 5

5 Further Research

6.1 Getting More Data

To improve model results, there are several methods which could be implemented. Gathering more data to train the model is the first and most obvious solution. Originally, 4 years of data from 2016- 2019 was used, however, in 2016 the variable NPS was not available and NPS turned out to be a significant contributor to the model. Therefore, the dataset was subset to only three years. It would be interesting to retest the 4 years of data without NPS, to see if the accuracy would improve.

6.2 Further Data Analysis

For further analysis, more time could be spent on data engineering and exploring the relationships between the response variables and independent variables. For example, “Not worth to switch” is one of the most important factors in driving customers decision to stay with an insurer for more than 6 years. It would be interesting to identify what the key drivers behind the variable “not worth to switch”. By identifying the key drivers of “Not worth to switch”, we may find more independent variables that would help improve the model accuracy.

6. Reference

[1] Zelterman, Daniel, *Applied Multivariate Statistics with R (Statistics for Biology and Health), 2015,* Chapter 8, pp. 208-211.

[2] Kassambara, Alboukadel *Machine Learning Essentials: Practical Guide in R*, 1ST edition, pp. 122- 127

[3] Brownlee, Jason, *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. Machine Learning Mastery, 21 Sept. 2016,

machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm- machine-learning/

[4] Liberman, Neil, *Decision Trees and Random Forests*, Towards Data Science, 26 Jan. 2017,

towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991/

[5] James, Gareth, Witten, Daniela, Hastie, Trevor and Tibshirani, Robert. *An Introduction to Statistical Learning: with Applications in R,* 7th edition, pp101-pp102

[6] *PCA: A Practical Guide to Principal Component Analysis in R & Python*, Analytics Vidhya, 21 March. 2016,

analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/

[7] James, Gareth, Witten, Daniela, Hastie, Trevor and Tibshirani, Robert. *An Introduction to Statistical Learning: with Applications in R.* 7th edition, Chapter 4

[8] Thakur, Debjoy, *Different ways of variable reduction*, Step Up for Analytics

stepupanalytics.com/different-ways-of-variable-reduction/

[9] Mic, *Predicting creditability using logistic regression in R: cross validating the classifier (part 2),* R-bloggers, 15 September. 2015. r-bloggers.com/predicting-creditability-using-logistic-regression-in-r-cross-validating-the-classifier-part-2-2/

[10] Atmathew, *Evaluating Logistic Regression Models*, R-bloggers,17 August. 2015, r-bloggers.com/evaluating-logistic-regression-models/

[11] biostat.jhsph.edu/~iruczins/teaching/jf/ch10.pdf

[12] Asaithambi, Sudharsan, *Why, How and When to Scale your Features,* Medium, 3 December. 2017, medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e

\